크롤링 프로젝트 개인 보고서

최태영

분석방법:

1. 티셀파, 티처빌 사이트 크롤링

* 방법:

Selenium과 webdriver 패키지를 이용한 동적 크롤링

* 기준:

날짜 기준으로 2017년도까지의 데이터 크롤링

* 데이터셋:

크롤링한 데이터 내역은 강의명, 작성날짜, 별점(최대5점), 후기내용 등이 있다.

1. 크롤링 데이터셋 전처리

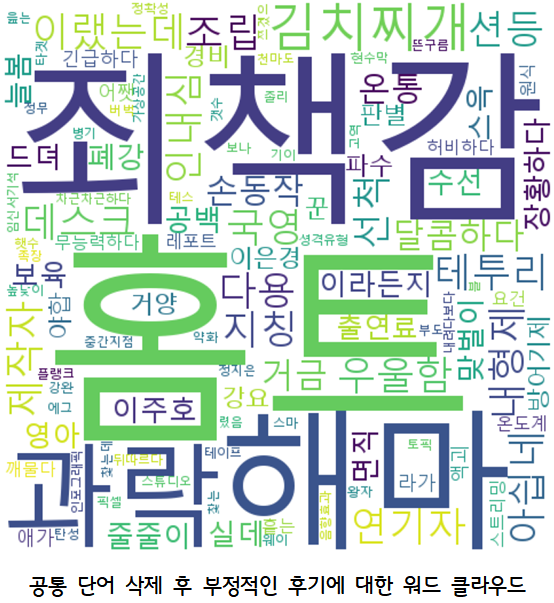
* 데이터셋에 각 컬럼명을 추가한뒤 별점 기준 4~5점을 label = 1, 1~3점을 label = 0으로 ‘label’ 컬럼을 추가한다.
* 완전 중복되는 샘플들을 제거한다.
* Null 값 유무 확인 후 제거한다.
* 날짜기준 최근 3개월내의 데이터만 남도록 데이터셋 편집한다.

1. . 자연어 처리

* Mecab 모델을 사용하여 morphs 함수를 이용해 후기내용을 전부 토큰화 한다. (‘tokenized’ 컬럼생성)
* 불용어 리스트를 이용하여 자연어 처리한다.
* 빈도수 높은 단어와 불필요성이 확인된 단어들을 추려서 drop list를 생성하여 토큰화 컬럼에 적용하여 제거한다.
* label값을 이용하여 긍정단어, 부정단어 집합 정리 후 서로 중복되는 단어들 리스트화 한다.
* 중복단어 리스트를 각 긍부정 단어집합에 적용, 제거한다.

1. 워드클라우드로 긍부정 단어 빈도수 시각화

* 티처빌 사이트 후기의 최근 3개월치 데이터 크롤링 후 별점 기준 라벨링을 진행하였다. 아이스크림 데이터셋과 동일하게 Okt 모델을 이용해 형태소 분석 뒤 긍부정 단어의 분포도를 확인하였다. 티처빌 데이터셋에서도 아이스크림과 동일하게 공통적으로 나타나는 단어들을 제거하고 부정적인 후기에서 빈도수가 높은 단어들을 키워드로 정의했다. 각 키워드를 통해 해당 후기들을 탐색, 분석하였다.
* 부정문 키워드



5. 높은 빈도의 부정문 가지고 해당 부정문이 들어가 있는 리뷰들 분석 정리

* 강의 내용:
  + 낮은 수준이다, 다양하지 않은 정보, 생각했던 내용이 아니다, 너무 기본적인 내용만 있다, 설명이 쓸데없이 길다, 강의 자료가 내용에 비해 부족하다, 세세한 자료 제공이 필요, 보조 도구 부족, 내용 최신화 필요, 강의 내용 실습화가 안된다, 컨텐츠 부족
* 강사평:
  + 강사가 산만하고 수다떠는 느낌
  + 강사 실력 저조
* 금액:
  + 강의 내용에 비해 비싸다
* 난이도:
  + 진도가 빠르다
  + 초보자용이 아니다
  + 자세한 설명 필요
  + 시험문제가 어렵다
  + 필수연수 치곤 어려운 난이도
* 강의 시스템
  + 자주 멈추고 끊기는 현상

강의 내용의 전반적인 퀄리티나 자료, 보조도구 보강 그리고 강의 컨텐츠의 최신화 및 업데이트의 필요성이 느껴진다. 또한 교원 대상 연수이다 보니 학생교육 시스템에 직접 접목 시킬수 있는 강의 내용들을 기대했지만 사례가 부족하다는 평이 많다. 따라서 연수 컨텐츠를 구성할 때 교육 시스템에 초점을 맞추는 것이 중요하다고 판단이 되었다. 마지막으로 전체적으로 강의 난이도가 높다는 의견이 많은 것으로 보아 시험문제나 진도속도 같은 강의 난이도 조절이 필요하다. 특히 필수연수들은 난이도를 쉽게 연수할 수 있도록 조정하는게 강의 인기도에 영향을 줄 것이라고 판단이 되었다.